УДК 004.89

СИГНАТУРНЫЕ МЕТОДЫ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

К. А. Мащенко[0000-0003-0355-6699]

НИЦ «Курчатовский институт» – НИИСИ, г. Москва, 117218, Россия

kirill.mashchenko@niisi.ru

Аннотация

Сигнатурные методы представляют собой мощный инструмент анализа временных рядов, который преобразует их в форму, удобную для задач машинного обучения. В статье рассмотрены основные понятия сигнатуры пути, ее свойства и геометрический смысл, а также методы вычисления для различных типов временных рядов. Приведены примеры применения сигнатурных методов в различных областях, включая финансы, медицину и образование, продемонстрированы их преимущества перед традиционными подходами. Особое внимание уделено генерации синтетических данных на основе сигнатур, что особенно актуально в условиях ограниченного объема исходных данных. Представлены результаты экспериментальных исследований по генерации и предсказанию траекторий цифрового следа обучения студентов, подтверждающие эффективность сигнатурных методов для применения в задачах машинного обучения по анализу и прогнозированию временных рядов.

Ключевые слова*: сигнатура, сигнатурные методы, временные ряды, генерация данных, анализ траекторий, цифровой след.*

ВВЕДЕНИЕ

Сигнатурные методы являются одним из эффективных инструментов для обработки и выделения признаков из пути – некоторый функции, аргументом которой является время, например, траектории случайного процесса. С помощью этого инструментария можно преобразовать временные ряды в форму, более подходящую для задач машинного обучения, таких как классификация и прогнозирование.

Понятие сигнатуры пути впервые было введено Ченом [1, 2], она представляет собой последовательность итерированных интегралов, которые отражают

различные свойства пути. Известно, что при выполнении некоторых достаточно общих условий, наложенных на рассматриваемый класс путей, существует взаимно однозначное соответствие между сигнатурами и путями. В рамках различных практических направлений, от задач восстановления траекторий до анализа потоков данных различной природы, было установлено, что достижение приемлемого уровня точности восстановления пути по его сигнатуре зачастую возможно при использовании лишь начальных членов сигнатуры, а именно первых нескольких повторных интегралов. Таким образом, сигнатура временного ряда, как способ кодирования, оказывается инструментом высокоэффективной компрессии информации, заключенной в траектории.

Временные ряды, являющиеся представлением реальных данных, по сути представляются своими значениями в дискретном множестве моментов времени. Таким образом, все рассматриваемые траектории являются кусочно-линейными, что упрощает определение сигнатуры и использование ее свойств.

В работах Чена, указанных выше, рассматривались достаточно регулярные траектории, то есть траектории с ограниченной вариацией, допускающие применение интегралов Римана–Стилтьеса. Для грубых траекторий (например, с показателем Гельдера меньше 1/2), которые требуют применения стохастических интегралов или интегралов по грубым траекториям, сигнатуры исследовались Лайонсом, Хайрером и др. [3–5]. Оказалось, что ряд результатов, например взаимнооднозначное соответствие между траекториями и сигнатурами, остается верным и для грубых траекторий [6, 7].

На протяжении последнего десятилетия наблюдается стремительное проникновение сигнатурных методов в арсенал прикладного машинного обучения. Так, будучи скомбинированными с архитектурами сверточных нейронных сетей, они обеспечили получение первенства в престижном онлайн-соревновании ICDAR 2013 по распознаванию изолированных китайских иероглифов [8], а в сопряжении с моделью градиентного бустинга заняли лидирующую позицию в состязании PhysioNet 2019, сосредоточенном на вычислительных аспектах кардиологической диагностики [9]. Кроме того, эти методы нашли широкое применение в области финансовой математики, в частности при решении задач хеджирования производных инструментов [10]. Для применения методов машинного обучения

в условиях недостаточного количества эмпирических данных была предложена методология генерации дополнительных траекторий, статистически согласованных с оригиналом [11].

Сигнатурные методы также успешно применяются в задачах классификации и регрессии. Например, в работе [12] сигнатуры использованы для анализа финансовых временных рядов, где они позволяют выделять ключевые особенности данных. Авторы установили, что даже небольшого количества коэффициентов сигнатуры достаточно для классификации финансовых потоков и прогнозирования рыночных изменений.

Традиционные методы анализа временных рядов, такие как динамические факторные модели (DFM) и авторегрессионные модели (ARIMA), часто требуют строгих предположений о стационарности и линейности данных. В отличие от них, сигнатурные методы не накладывают таких ограничений. В работе [13] показано, что регрессия на сигнатурах превосходит DFM в задаче прогнозирования ВВП, обеспечивая меньшую ошибку и большую устойчивость к нерегулярностям данных.

Одной из проблем сигнатурных методов является экспоненциальный рост числа сигнатурных членов с увеличением уровня усечения. Для решения этой проблемы применяются методы регуляризации, такие как LASSO и Elastic Net [14], а также методы уменьшения размерности, например метод главных компонент (PCA). В работе [15] обсуждается, как выбор уровня усечения и стандартизация элементов сигнатуры влияют на качество моделей.

МЕТОДЫ И МАТЕРИАЛЫ

Сигнатурные методы

С теоретической точки зрения сигнатурный подход следует квалифицировать как непараметрическую, устойчивую к шумам методику извлечения репрезентативных признаков, способных впоследствии служить входными признаками для моделей обучения. Его фундаментальная сила заключается в способности по параметризованному пути, задающему последовательность наблюдаемых со-

стояний, сформировать компактный, но исчерпывающий набор признаков, аккумулирующих как аналитические, так и геометрические характеристики рассматриваемого процесса.

Определение и свойства пути

Путь в $\mathrm{R^d}$ – это непрерывное отображение X из некоторого интервала [a, b] в R^d. Чтобы подчеркнуть зависимость от времени, используют обозначение $\mathrm{X_t} = \mathrm{X(t)}\colon [\mathrm{a}, \mathrm{b}] \to \mathrm{R^d}$.

В дальнейшем будем предполагать, что рассматриваемые пути являются достаточно «хорошими» отображениями, а именно, они являются кусочно-дифференцируемыми (вообще говоря, справедливость большинства результатов сохранится, если считать, что пути имеют ограниченную вариацию). Будем называть путь гладким, если он бесконечно дифференцируем.

Интегралом от функции $f\colon R \mapsto R$ по одномерному пути $X\colon [a, b] \mapsto R$ называется величина

$$\int_a^b f(X_t)dX_t = \int_a^b f(X_t)\dot{X}_t\, dt,$$

где последний интеграл является обычным (римановым) интегралом непрерывной ограниченной функции. Обозначение «верхняя точка» здесь и далее использовано для дифференцирования по одной переменной: $\dot{X}_t = dX_t/dt$.

Заметим, что $f(X_t)$ тоже является путем на $[a, b]$, поэтому можно естественным образом определить интеграл от пути по пути.

Интегралом от пути $Y\colon [a, b] \mapsto R$ по пути $X\colon [a, b] \mapsto R$ называется величина

$$\int_a^b Y_t dX_t = \int_a^b Y_t\, \dot{X}_t dt.$$

Определение сигнатуры пути

Обозначим координаты пути $X\colon [a, b] \mapsto R^d$ как $\left(X_t^1, \dots, X_t^d\right)$, где каждая координата $X^i\colon [a, b] \mapsto R$ является путем. Для всех $i \in \{1, \dots, d\}$ и $t \in [a, b]$ определим величину

$$S(X)_{a,t}^{i} = \int_{a<s<t} d\,X_s^i = X_t^i - X_a^i,$$

которая является приращением *i*-й координаты пути до момента времени *t*. Отметим, что $S(X)_{a,\cdot}^{i}$ – это тоже путь.

Теперь определим для любой пары $i, j \in \{1, \dots, d\}$ двойной повторный интеграл

$$S(X)_{a,t}^{i,j} = \int_{a<s<t} S(X)_{a,s}^{i}\, dX_s^j = \int_{a<r<s<t} d\,X_i^r\, dX_s^j = \int_{a<s<t}\int_{a<r<s} d\,X_i^r\, dX_s^j.$$

Можно продолжить по индукции: для любых $k \geq 1$ и набора индексов $i_1, \dots, i_k \in \{1, \dots, d\}$ определим

$$S(X)_{a,t}^{i_1,\dots,i_k} = \int_{a<s<t} S(X)_{a,s}^{i_1,\dots,i_{k-1}} \mathrm{d} X_s^{i_k} = \int_{a<t_k<t} \dots \int_{a<t_1<t_2} d\,X_{t_1}^{i_1} \dots dX_{t_k}^{i_k}.$$

Величина $S(X)_{a,t}^{i_1,\dots,i_k}$ называется *k*-кратным повторным интегралом от пути *X* по индексам $i_1, \dots, i_k$.

Сигнатурой пути $X\colon [a,b] \mapsto R^d$ называется бесконечный набор $S(X)_{a,b}$ всех повторных интегралов от X:

$$S(X)_{a,b} = \left(1, S(X)_{a,b}^{1}, \dots, S(X)_{a,b}^{d}, S(X)_{a,b}^{1,1}, S(X)_{a,b}^{1,2}, \dots\right),$$

где первый элемент сигнатуры (соответствующий пустому индексу) по определению считается равным 1, а верхние индексы остальных элементов пробегают набор всевозможных мульти-индексов

$$W = \{(i_1, \dots, i_k) \mid k \geq 1, \qquad i_1, \dots, i_k \in \{1, \dots, d\}\}.$$

Множество *W* называется множеством слов в алфавите $A = \{1, \dots, d\}$. Конечный набор чисел $S(X)_{a,b}^{i_1,\dots,i_k}$ для всевозможных мультииндексов длины *k* будем называть *k*-м уровнем сигнатуры.

Пример 1. Рассмотрим произвольный одномерный путь $X\colon [a,b] \mapsto R$. Тогда сигнатура этого пути вычисляется следующим образом:

$S(X)_{a,b}^{1} = X_b - X_a,$

$\mathrm{S}(X)_{a,b}^{1,1} = \frac{(X_b - X_a)^2}{2!},$

$\mathrm{S}(X)_{a,b}^{1,1,1} = \frac{(X_b - X_a)^3}{3!},$

...

$\mathrm{S}(X)_{a,b}^{1,1,\ldots,1} = \frac{(X_b - X_a)^k}{k!}$.

Отсюда можно видеть, что сигнатура по повторным индексам выражается через приращение по соответствующей координате, что верно и для многомерного случая.

Пример 2. Рассмотрим следующий двумерный путь X: $[a, b] \mapsto R^2$:

$$X_t = \{X_t^1, X_t^2\} = \{t, k \cdot t + d\},$$

$$dX_t = \{dX_t^1, dX_t^2\} = \{dt, k \cdot dt\},$$

где $k \neq 0$ – параметр. Тогда элементы сигнатуры $S(X)_{a,b}^{1,2}$ и $S(X)_{a,b}^{2,1}$ этого пути вычисляются следующим образом:

$$S(X)_{a,b}^{1,2} = \int_a^b \left(\int_a^{t_2} dt_1 \right) k\, dt_2 = \frac{k \cdot (b-a)^2}{2},$$

$$S(X)_{a,b}^{2,1} = \int_a^b \left(\int_a^{t_2} k\, dt_1 \right) dt_2 = \frac{k \cdot (b-a)^2}{2}.$$

Шафл-произведение

Одной из важных алгебраических особенностей сигнатуры является то, что произведение двух ее элементов $S(X)_{a,b}^{i_1,\ldots,i_k}$ и $S(X)_{a,b}^{j_1,\ldots,j_m}$ всегда может быть выражено через сумму других ее элементов, которая зависит исключительно от мультииндексов $(i_1, \ldots, i_k)$ и $(j_1, \ldots, j_m)$. Это свойство влечет за собой важнейшее следствие, показывающее отсутствие алгебраической независимости между членами сигнатуры. Кроме того, оно позволяет отказаться от непосредственного манипулирования произведениями элементов сигнатуры и перейти к работе с линейными комбинациями, что, в свою очередь, существенно упрощает их аналитическую обработку.

Далее введем понятие шафл-произведения двух мультииндексов.

Определение. *Перестановка* σ *множества* $\{1, \ldots, k+m\}$ *называется* ***(k,m)-шафлом****, если* $\sigma^{-1}(1) < \cdots < \sigma^{-1}(k)$ и $\sigma^{-1}(k+1) < \cdots < \sigma^{-1}(k+m)$.

Для множества всех (*k,m*)*-шафлов будем использовать обозначение* $\mathrm{Shuffles}(k,m)$.

Рассмотрим два мульти-индекса $\mathrm{I} = (i_1, \dots, i_k)$ и $J = (j_1, \dots, j_m)$, $i_1, \dots, i_k, j_1, \dots, j_m \in \{1, \dots, d\}$. Определим мультииндекс

$$(r_1, \dots, r_k, r_{k+1}, \dots, r_{k+m}) = (i_1, \dots, i_k, j_1, \dots, j_m).$$

Шафл-произведением *I* и *J* (обозначение: $I \text{ш} J$) называется конечный набор мульти-индексов длины *k*+*m* вида

$$I \text{ш} J \; = \; \{(r\_\{\sigma(1)\}, \dots, r\_\{\sigma(k+m)\} \mid \sigma \in \mathrm{Shuffles}(k,m)\}.$$

Теорема. *Для любого пути* $X\colon [a,b] \mapsto R^d$ *и мультииндексов* $I = (i_1, \dots, i_k)$ *и* $J = (j_1, \dots, j_m)$, $i_1, \dots, i_k, j_1, \dots, j_m \in \{1, \dots, d\}$ *верно равенство*

$$S(X)^I_{a,b} S(X)^J_{a,b} = \sum_{K \in I \text{ш} J} S(X)^K_{a,b}.$$

Пример 3. Произведение элементов сигнатуры первого порядка выражается через элементы второго порядка следующим образом:

$$S(X)^1_{a,b} S(X)^2_{a,b} = S(X)^{1,2}_{a,b} + S(X)^{2,1}_{a,b}.$$

Геометрический смысл сигнатур

Независимость от начальной точки. Рассмотрим путь $X\colon [a,b] \mapsto R^d$ и $h \in R^d$. Пусть путь $Y\colon [x,y] \mapsto R^d$ имеет вид $Y_t = X_t + h$. Тогда

$$S(X)_{a,b} = S(Y)_{a,b}.$$

Независимость от репараметризации времени. Рассмотрим путь $X\colon [a,b] \mapsto R^d$ и биективную непрерывную неубывающую функцию $\psi\colon [x,y] \mapsto [a,b]$. Пусть путь $Y\colon [x,y] \mapsto R^d$ имеет вид $Y_t = X_{\psi_t}$. Тогда

$$S(X)_{a,b} = S(Y)_{x,y}.$$

Отметим, что первый уровень сигнатуры $\left(S(X)^1_{a,b}, \dots, S(X)^d_{a,b}\right)$, как видно из ее формального определения, совпадает с приращениями пути по отдельным координатам, в то время как второй уровень интерпретируется через понятие площади Леви (см., например, [3, 10]) – геометрической характеристики, указывающей на ориентированную площадь, охватываемую траекторией в двумерном пространстве.

Пусть $X\colon [a,b] \mapsto R^2$ – двумерный путь, где $X_t = (X_t^1, X_t^2)$. Проведем прямую от начальной до конечной точки пути, после чего все получившиеся площади при пересечении этой прямой и пути рассмотрим со знаком минус, если они выше прямой, и со знаком плюс, если они ниже прямой. **Площадью Леви** называется сумма данных площадей с учетом знаков, см. рис. 1.

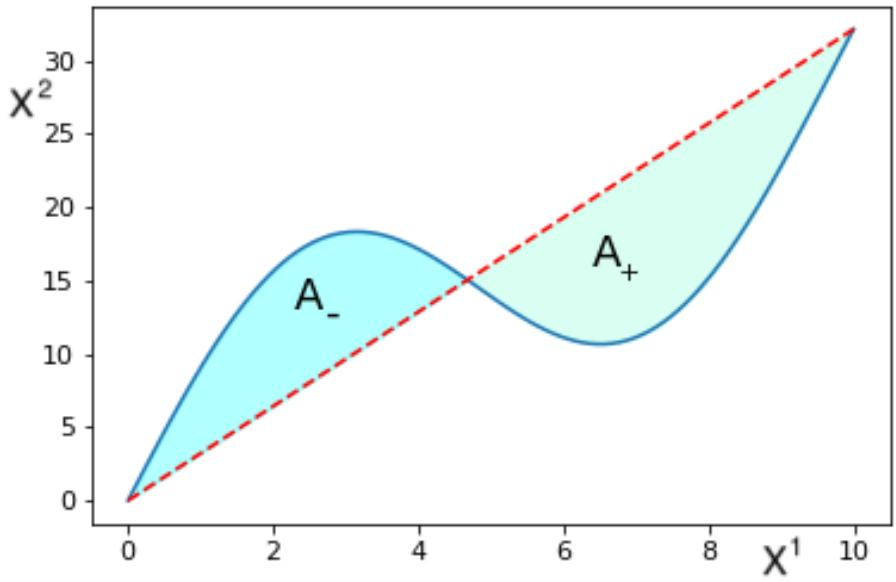

Рис. 1. Площадь Леви $A=A_+A_-$

Площадь Леви выражается через элементы сигнатуры следующим образом:

$$A = \frac{1}{2}\left(S(X)_{a,b}^{1,2} - S(X)_{a,b}^{2,1}\right).$$

Тождество Чена

Эффективность вычисления сигнатур в значительной мере зависит от природы исходного пути. В тех случаях, когда имеется аналитически заданная параметризация, соответствующие повторные интегралы могут быть получены в явной форме. Однако в большинстве прикладных задач, где траектории представляют собой последовательности наблюдений, их форма носит кусочно-линейный характер, а параметризация затруднена или вовсе отсутствует. В подобных случаях оказывается целесообразным применение тождества Чена, позволяющего рекурсивно вычислять сигнатуру всего пути по сигнатурам его составных отрезков.

Тождество Чена. Пусть $a < b < c$ и $X: [a, c] \mapsto R^d$. Тогда для любых $i_1, \dots, i_k \in W$ выполнено равенство

$$S(X)_{a,c}^{i_1,\dots,i_k} = \sum_{m=0}^{k} S(X)_{a,b}^{i_1,\dots,i_m} S(X)_{b,c}^{i_{m+1},\dots,i_k}.$$

Для каждого одномерного сегмента сигнатура может быть получена напрямую, после чего производится поочередное соединение с предыдущими с использованием вышеуказанного тождества.

Пример 4. Рассмотрим путь $X: [0,2] \mapsto R^2$ следующего вида, рис. 2:

$$X_t = \{t, 2 \cdot t\}, \qquad t \in [0,1],$$

$$X_t = \{t, 3 - t\}, \qquad t \in [1,2].$$

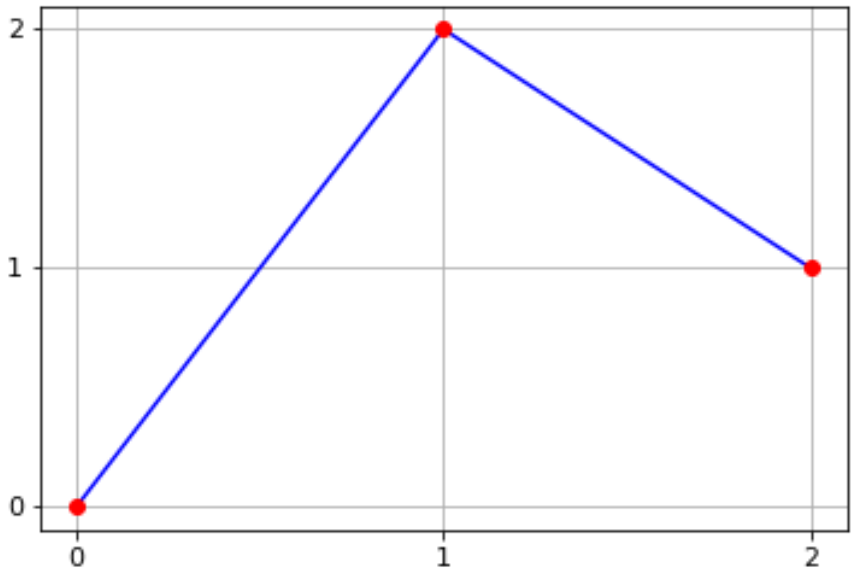

Рис. 2. Пример кусочно-линейного пути

Найдем первые два уровня сигнатуры пути отдельно на каждом из отрезков [0,1], [1,2]. На этих отрезках путь является одномерным и линейным, поэтому, согласно примерам 1 и 2, первые два уровня сигнатуры этих участков выражаются следующим образом.

На отрезке [0,1]:

$S(X)_{0,1}^{\emptyset} = 1$ (по определению),

$S(X)_{0,1}^{1} = \Delta X_{0,1}^{1} = 1,$

$S(X)_{0,1}^{1,1} = \frac{\left(\Delta X_{0,1}^{1}\right)^2}{2} = 0.5,$

$S(X)_{0,1}^{2} = \Delta X_{0,1}^{2} = 2,$

$S(X)_{0,1}^{2,2} = \frac{\left(\Delta X_{0,1}^{2}\right)^2}{2} = 2,$

$S(X)_{0,1}^{1,2} = \frac{2 \cdot (1-0)^2}{2} = 1,$

$S(X)_{0,1}^{2,1} = \frac{2 \cdot (1-0)^2}{2} = 1.$

На отрезке [1,2]:

$S(X)_{1,2}^{\emptyset} = 1$ (по определению),

$S(X)_{1,2}^{1} = \Delta X_{1,2}^{1} = 1,$

$S(X)_{1,2}^{1,1} = \frac{\left(\Delta X_{1,2}^{1}\right)^2}{2} = 0.5,$

$S(X)_{1,2}^{2} = \Delta X_{1,2}^{2} = -1,$

$S(X)_{1,2}^{2,2} = \frac{\left(\Delta X_{1,2}^{2}\right)^2}{2} = 0.5,$

$S(X)_{1,2}^{1,2} = \frac{-1 \cdot (2-1)^2}{2} = -0.5,$

$S(X)_{1,2}^{2,1} = \frac{-1 \cdot (2-1)^2}{2} = -0.5.$

Применим тождество Чена:

$S(X)_{0,2}^{1} = S(X)_{0,1}^{1} \cdot S(X)_{1,2}^{\emptyset} + S(X)_{0,1}^{\emptyset} \cdot S(X)_{1,2}^{1} = 1 \cdot 1 + 1 \cdot 1 = 2 = \Delta X_{0,2}^{1},$

$S(X)_{0,2}^{2} = S(X)_{0,1}^{2} \cdot S(X)_{1,2}^{\emptyset} + S(X)_{0,1}^{\emptyset} \cdot S(X)_{1,2}^{2} = 2 \cdot 1 + 1 \cdot (-1) = 1 = \Delta X_{0,2}^{2},$

$S(X)_{0,2}^{1,2} = S(X)_{0,1}^{1,2} \cdot S(X)_{1,2}^{\emptyset} + S(X)_{0,1}^{1} \cdot S(X)_{1,2}^{2} + S(X)_{0,1}^{\emptyset} \cdot S(X)_{1,2}^{1,2}$

$= 1 \cdot 1 + 1 \cdot (-1) + 1 \cdot (-0.5) = -0.5,$

$S(X)_{0,2}^{2,1} = S(X)_{0,1}^{2,1} + S(X)_{0,1}^{2} \cdot S(X)_{1,2}^{1} + S(X)_{1,2}^{2,1} = 1 + 2 \cdot 1 - 0.5 = 2.5,$

$$S(X)_{0,2}^{1,1} = S(X)_{0,1}^{1,1} + S(X)_{0,1}^{1} \cdot S(X)_{1,2}^{1} + S(X)_{1,2}^{1,1} = 0.5 + 1 \cdot 1 + 0.5 = 2 = \frac{\left(\Delta X_{0,2}^{1}\right)^2}{2},$$

$$S(X)_{0,2}^{2,2} = S(X)_{0,1}^{2,2} + S(X)_{0,1}^{2} \cdot S(X)_{1,2}^{2} + S(X)_{1,2}^{2,2} = 2 + 2 \cdot (-1) + 0.5 = 0.5 = \frac{\left(\Delta X_{0,2}^{2}\right)^2}{2}.$$

Логарифмические сигнатуры

Из теоремы о шафл-произведении вытекает, что элементы сигнатуры не являются алгебраически независимыми, однако для многих моделей машинного обучения требуются независимые входные данные для корректной работы и лучшего качества. Эту проблему решает понятие логарифмических сигнатур, которые, в отличие от обычных сигнатур, представляют собой алгебраически независимый набор величин, подходящих для прямого использования в качестве признакового пространства моделей.

Для определения понятия логарифмической сигнатуры введем сначала алгебру формальных степенных рядов.

Рассмотрим *d* формальных неопределенных величин $e_1, \dots, e_d$. Алгеброй некоммутативных формальных степенных рядов из *d* неопределенных называется векторное пространство всех рядов вида

$$\sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k=1}^{d} \lambda_{i_1,\dots,i_k} e_{i_1} \dots e_{i_k},$$

где параметр второй суммы пробегает по всем мультииндексам $(i_1, \dots, i_k)$, $i_1, \dots, i_k \in 1, \dots, d$, и коэффициенты $\lambda_{i_1,\dots,i_k}$ являются вещественными числами.

На этом пространстве определены стандартные операции сложения рядов и умножения ряда на коэффициент, а также операция умножения рядов $\otimes$. При этом элементы пространства некоммутативны, т. е., например, $e_1 e_2$ и $e_2 e_1$ являются разными элементами.

Сигнатура может быть «закодирована» как элемент этого пространства (т. е. существует взаимно однозначное соответствие между сигнатурами и формальными степенными рядами):

$$S(X)_{a,b} = \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k=1}^{d} S(X)_{a,b}^{i_1,\dots,i_k} e_{i_1} \dots e_{i_k}.$$

Иными словами, элементы сигнатуры рассматриваются как соответствующие коэффициенты ряда, что позволяет «перейти» в алгебру рядов. В ней можно выполнять операции сложения и умножения рядов, а коэффициенты полученного ряда интерпретировать как элементы сигнатуры.

Далее для формального ряда, в котором первый коэффициент равен 1, а остальные равны 0 (что соответствует сигнатуре $\{1,0,0,\dots\}$), будем использовать обозначение 1.

Логарифмической сигнатурой пути $X\colon [a,b] \mapsto R^d$ называется формальный степенной ряд

$$\log S\,(X)_{a,b} = \sum_{n=1} \frac{(-1)^n}{n} \left(1 - S(X)_{a,b}\right)^{\otimes n}.$$

Тем самым, логарифмическая сигнатура становится инструментом генерации независимого набора признаков, обладающих полной описательной способностью по отношению к геометрии исходного многомерного пути.

Следует, однако, указать, что сигнатура не позволяет восстановить траекторию в полном объеме. Так, например, информация о скорости прохождения пути утрачивается в силу инвариантности сигнатуры относительно репараметризации времени. Тем не менее при соблюдении определенных условий, в частности при отсутствии самопересечений, сигнатура полностью определяет как множество точек, через которые проходит путь, так и порядок их обхода, что открывает возможности для применения в задачах восстановления структуры данных.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Практика предоставляет множество различных задач, в которых неизвестны распределение и стохастическая основа данных, в связи с чем нет возможности установить математические свойства модели. Модели, описывающие эти задачи, можно значительно усложнять, подбирая и добавляя в них новые параметры, однако в реальности они не смогут полностью описать наблюдаемое поведение.

Помимо этого, при использовании временных рядов во многих ситуациях может оказаться недостаточным количество данных для применения моделей глубокого машинного обучения, а также данные могут быть конфиденциальны и недоступны для обучения моделей. Примерами этого могут быть образовательные, медицинские и финансовые данные.

В ситуациях, когда необходимо расширить доступное множество данных, не прибегая к гипотезам об их вероятностной природе, можно использовать сигнатурные методы в сочетании с генеративными моделями. Так, на основе оригинального множества временных рядов, представленных в сигнатурной форме, можно синтезировать новые данные, демонстрирующие похожее распределение. При этом необходимость знания истинного распределения отпадает, поскольку оценка схожести может быть проведена с помощью, например, метрики максимального среднего расхождения (MMD), построенной также на основе сигнатур [16].

Данные образовательного цифрового следа студентов являются ярким примером временных рядов для которых неизвестны их распределение и стохастическая основа. Кроме того, в зависимости от университета, курса и среднего уровня студентов они могут отличаться, что вводит еще большие ограничения на количество доступных данных. Поэтому для тестирования различных педагогических гипотез, а также для создания моделей по подбору персональных траекторий и построению поведенческой и предсказательной аналитики необходимы многолетние сборы студенческих данных.

В связи с этим были применены методы генерации, анализа и прогноза траекторий обучения студентов, использующие сигнатуры исходного набора данных. Преимуществом использования сигнатурных методов является то, что они способны улавливать взаимосвязи между различными размерностями временного ряда. Например, появляется возможность обнаруживать взаимные зависимости между посещаемостью, временем, потраченным на решение задач, плагиатом, долей набранных за занятия баллов и другими метриками цифрового следа студентов. Это значительно повышает общее качество модели, поскольку рассмотрение отдельных характеристик независимо друг от друга не дает полной картины, например, у студента могут быть высокие баллы, но при высоком уровне

плагиата, что изменяет его общий статус относительно группы. Помимо этого, у модели появляется возможность улавливать, например, повышение плагиата при сокращении времени решения задач, снижение среднего балла при падении посещаемости и другие зависимости.

Такой подход, реализованный в контексте образовательной аналитики, позволяет на массиве данных, синтетически сгенерированных с помощью сигнатурных методов, построить модель прогнозирования поведения студента, в которой сигнатуры временных рядов выступают в роли признакового пространства. Эта модель, в свою очередь, служит инструментом визуализации и интерпретации динамики учебных результатов, указывая преподавателю на необходимость индивидуального вмешательства или корректировки образовательной траектории с целью максимизации обучающего эффекта. Более того, на основе такого прогностического механизма можно автоматически сформировать персонализированные образовательные траектории, предложив учащимся релевантные задания, а также вспомогательные материалы, соответствующие их текущему уровню и способствующие увеличению эффективности их обучения.

На рис. 3, представлены результаты работы обученной прогностической модели, которая предсказывает следующее значение траектории цифрового следа студента на основе окна из 5 предыдущих значений. Вертикальной линией отмечен момент, с которого модель начинает свою работу, таким образом, уже спустя примерно месяц реального преподавания у модели появляется возможность предсказывать будущее поведение студента. При сдвиге окна каждый раз в качестве нового значения добавляется истинное значение. Такой подход хорошо соотносится с реальным педагогическим применением модели, так как она в режиме реального времени имеет доступ к актуальным данным студента на текущую дату.

Рис. 3. Пример работы модели предсказания траектории студента с помощью сигнатурных методов.

ЗАКЛЮЧЕНИЕ

Сигнатурные методы доказали свою эффективность при анализе временных рядов, предложив гибкий инструмент для машинного обучения и генерации данных. Они преодолевают ограничения традиционных подходов и показывают высокую точность в прогнозировании. Особую ценность сигнатурные методы представляют в условиях ограниченного объема данных, поскольку позволяют генерировать синтетические данные, сохраняющие распределение исходных. Это открывает новые возможности для применения в образовании, медицине, финансах и других областях, где данные часто являются конфиденциальными или труднодоступными. Результаты экспериментальных исследований, представленные в статье, подтверждают практическую значимость сигнатурных методов, особенно в анализе образовательных траекторий студентов. Использование этих методов позволяет не только прогнозировать поведение студентов, но и адаптировать учебные программы для достижения наилучших результатов.

Благодарности

Работа выполнена в рамках темы государственного задания НИЦ «Курчатовский институт» – НИИСИ по теме № FNEF-2024-0001 (1023032100070-3-1.2.1).

СПИСОК ЛИТЕРАТУРЫ

1. *Chen K.-T.* Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula // Annals of Mathematics. 1957. Vol. 65. No. 1. P. 163–178. https://doi.org/10.2307/1969671

2. *Chen K.-T.* Integration of paths – a faithful representation of paths by noncommutative formal power series // Transactions of the American Mathematical Society. 1958. Vol. 89. P. 395–407. https://doi.org/10.1090/S0002-9947-1958-0106258-0

3. *Lyons T.J.* Differential equations driven by rough signals // Revista Matematica Iberoamericana. 1998. Vol. 14. No. 2. P. 215–310.
https://doi.org/10.4171/RMI/240

4. *Lyons T.J., Caruana M., Levy T.* Differential equations driven by rough paths // Lecture Notes in Mathematics. 2007. № 1908.
https://doi.org/10.1007/978-3-540-71285-5

5. *Friz P.K., Hairer M.* A course on rough paths. With an introduction to regularity structures (2nd edition) // Springer. 2020.
https://doi.org/10.1007/978-3-030-41556-3

6. *Boedihardjo H., Geng X., Lyons T., Yang D.* The signature of a rough path: Uniqueness. 2014. https://arxiv.org/abs/1406.7871

7. *Hambly B., Lyons T.* Uniqueness for the signature of a path of bounded variation and the reduced path group // Annals of Mathematics. 2010. Vol. 171. No. 1. P. 109–167. https://doi.org/10.4007/annals.2010.171.109

8. *Graham B.* Sparse arrays of signatures for online character recognition. 2013. https://arxiv.org/abs/1308.0371

9. *Morrill J., Kormilitzin A., Nevado-Holgado A., Swaminathan S., Howison S., Lyons T.J.* The signature-based model for early detection of sepsis from electronic health records in the intensive care unit // IEEE Conference on Computing in Cardiology. 2019. https://doi.org/10.22489/CinC.2019.014

10. *Chevyrev I., Kormilitzin A.* A Primer on the Signature Method in Machine Learning. 2019. https://arxiv.org/abs/1603.03788

11. *Buhler H., Horvath B., Lyons T., Arribas I. P., Wood B.* A data-driven market simulator for small data environments. 2020. https://arxiv.org/abs/2006.14498

12. *Gyurko L.G., Lyons T., Kontkowski M., Field J.* Extracting information from the signature of a financial data stream. 2014. https://arxiv.org/abs/1307.7244v2

13. *Cohen S.N., Lui S., Malpass W., Mantoan G., Nesheim L., de Paula A., Reeves A., Scott C., Small E., Yang L.* Nowcasting with signature methods. 2023. https://arxiv.org/abs/2305.10256v1

14. *Zou H., Hastie T.* Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005. Vol. 67. No. 2. P. 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

15. *Fermanian A.* Embedding and learning with signatures // Computational Statistics and Data Analysis. 2021. Vol. 157. No. 107148. https://doi.org/10.1016/j.csda.2020.107148

16. *Chevyrev I., Oberhauser H.* Signature moments to characterize laws of stochastic processes. 2018. https://arxiv.org/abs/1810.10971

SIGNATURE METHODS FOR TIME SERIES ANALYSIS

K. A. Mashchenko[0000-0003-0355-6699]

NRC "Kurchatov Institute" – SRISA, Moscow, 117218, Russia

kirill.mashchenko@niisi.ru

Abstract

Signature methods are a powerful tool for time series analysis, transforming them into a form suitable for machine learning tasks. The article examines the fundamental concepts of path signatures, their properties, and geometric interpretation, as well as computational methods for various types of time series. Examples of signature method applications in different fields, including finance, medicine, and education, are presented, highlighting their advantages over traditional approaches. Special attention is given to synthetic data generation based on signatures, which is particularly relevant

when working with limited datasets. The experimental results on generating and predicting student digital learning trajectories demonstrate the effectiveness of signature methods for machine learning applications in time series analysis and forecasting.

Keywords*: signature, signature methods, time series, data generation, trajectory analysis, digital footprint*

REFERENCES

1. *Chen K.-T.* Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula // Annals of Mathematics. 1957. Vol. 65. No. 1. P. 163–178. https://doi.org/10.2307/1969671

2. *Chen K.-T.* Integration of paths – a faithful representation of paths by noncommutative formal power series // Transactions of the American Mathematical Society. 1958. Vol. 89. P. 395–407. https://doi.org/10.1090/S0002-9947-1958-0106258-0

3. *Lyons T.J.* Differential equations driven by rough signals // Revista Matematica Iberoamericana. 1998. Vol. 14. No. 2. P. 215–310.
https://doi.org/10.4171/RMI/240

4. *Lyons T.J., Caruana M., Levy T.* Differential equations driven by rough paths // Lecture Notes in Mathematics. 2007. № 1908.
https://doi.org/10.1007/978-3-540-71285-5

5. *Friz P.K., Hairer M.* A course on rough paths. With an introduction to regularity structures (2nd edition) // Springer. 2020.
https://doi.org/10.1007/978-3-030-41556-3

6. *Boedihardjo H., Geng X., Lyons T., Yang D.* The signature of a rough path: Uniqueness. 2014. https://arxiv.org/abs/1406.7871

7. *Hambly B., Lyons T.* Uniqueness for the signature of a path of bounded variation and the reduced path group // Annals of Mathematics. 2010. Vol. 171. No. 1. P. 109–167. https://doi.org/10.4007/annals.2010.171.109

8. *Graham B.* Sparse arrays of signatures for online character recognition. 2013. https://arxiv.org/abs/1308.0371

9. *Morrill J., Kormilitzin A., Nevado-Holgado A., Swaminathan S., Howison S., Lyons T.J.* The signature-based model for early detection of sepsis from electronic health records in the intensive care unit // IEEE Conference on Computing in Cardiology. 2019. https://doi.org/10.22489/CinC.2019.014

10. *Chevyrev I., Kormilitzin A.* A Primer on the Signature Method in Machine Learning. 2019. https://arxiv.org/abs/1603.03788

11. *Buhler H., Horvath B., Lyons T., Arribas I. P., Wood B.* A data-driven market simulator for small data environments. 2020. https://arxiv.org/abs/2006.14498

12. *Gyurko L.G., Lyons T., Kontkowski M., Field J.* Extracting information from the signature of a financial data stream. 2014. https://arxiv.org/abs/1307.7244v2

13. *Cohen S.N., Lui S., Malpass W., Mantoan G., Nesheim L., de Paula A., Reeves A., Scott C., Small E., Yang L.* Nowcasting with signature methods. 2023. https://arxiv.org/abs/2305.10256v1

14. *Zou H., Hastie T.* Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005. Vol. 67. No. 2. P. 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

15. *Fermanian A.* Embedding and learning with signatures // Computational Statistics and Data Analysis. 2021. Vol. 157. No. 107148. https://doi.org/10.1016/j.csda.2020.107148

16. *Chevyrev I., Oberhauser H.* Signature moments to characterize laws of stochastic processes. 2018. https://arxiv.org/abs/1810.10971

СВЕДЕНИЯ ОБ АВТОРЕ

МАЩЕНКО Кирилл Алексеевич. Младший научный сотрудник отдела учебной информатики НИЦ «Курчатовский институт» – НИИСИ, младший научный сотрудник Лаборатории вычислительных методов механико-математического факультета МГУ имени М.В. Ломоносова. Области исследований: автоматизация и цифровизация образовательного процесса, в том числе с применением методов искусственного интеллекта.

Kirill A. MASHCHENKO. Junior Researcher at the Department of Educational Informatics, NRC “Kurchatov Institute” – SRISA, Junior Researcher at the Laboratory of Computational Methods, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University. Research areas: Automation and digital transformation of the educational process, including the application of artificial intelligence methods.

email: kirill.mashchenko@niisi.ru

ORCID: 0000-0003-0355-6699

Материал поступил в редакцию 10 апреля 2025 года